vsTECNIRIS-33

Tecnologías de Almacenamiento (2ª Parte)

Red IRIS

Octubre 2014
Gabriel Verdejo Álvarez

ÍNDICE

Presentación disponible en http://rdlab.cs.upc.edu/docu/vstecniris33.pdf

RDlab – CS, UPC

- ***Sobre el RDlab***

  1986 Creación del departamento de Lenguajes y Sistemas Informáticos (LSI)

  1987 Creación del Laboratorio de Cálculo de LSI (LCLSI)

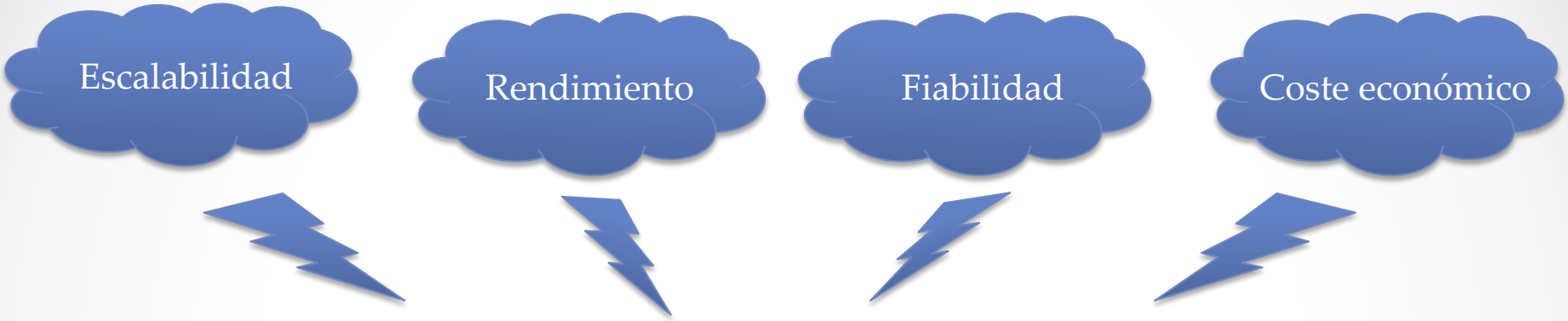
  2010 Creación del Laboratorio de investigación y desarrollo en LSI (RDlab)

  2014 Cambio de nombre departamental: Ciencias de la computación (CS)

- ***RDlab en números***

  - Trabajamos con más de 10 grupos de investigación y centros afines
  - Proporcionamos soporte a más de 150 investigadores y PDI
  - Gestionamos más de 150 servidores, 1000 cores y 120 TBytes de disco

Sistemas de ficheros de alto rendimiento



Sistemas de ficheros de alto rendimiento II

- ***Características principales***

  - Multiservidor

  - Separación de datos y metadatos

  - Escalable

- ***Extras***

  - Redundancia / Tolerancia a fallos

  - Balanceo de la carga de peticiones

  - Integración en aplicativos

Sistemas de ficheros de alto rendimiento III

- ***Inicio de los HPFS***

  - Inicialmente en entornos de supercomputación / experimentales

  - Software privativo

- ***Popularización de los HPFS***

  - Abaratamiento del hardware

  - Linux

  - Alternativas libres

  - Madurez de los proyectos

Nuestra experiencia

- ***(breve) Historia del uso de HPFS en el RDlab***

  - Experiencia desde el 2005

  - Entornos de supercomputación

  - Hardware no dedicado (inicialmente ;)

  - GlusterFS (2 años en producción) y Lustre (más de 4 años en producción)

- ***Análisis HPFS en el 2014***

  - Otros entornos de uso para HPFS (*OpenNebula*)

  - El soporte a la investigación ("*estar a la última*" vs conocer lo último)

  - Lustre vs Gluster vs BeeGFS (FhGFS)

Nuestra experiencia II

- ***Premisas de la evaluación***

  - Aumento del rendimiento

  - "Respeto" a la arquitectura existente:

    1. Conexiones 1 Gbit para clientes y 10Gbit Ethernet para servidores de datos
    2. Arrays de disco con conexión "*Fiber Channel*" dedicados (Fujitsu)
    3. Clientes nativos para sistemas GNU/Linux

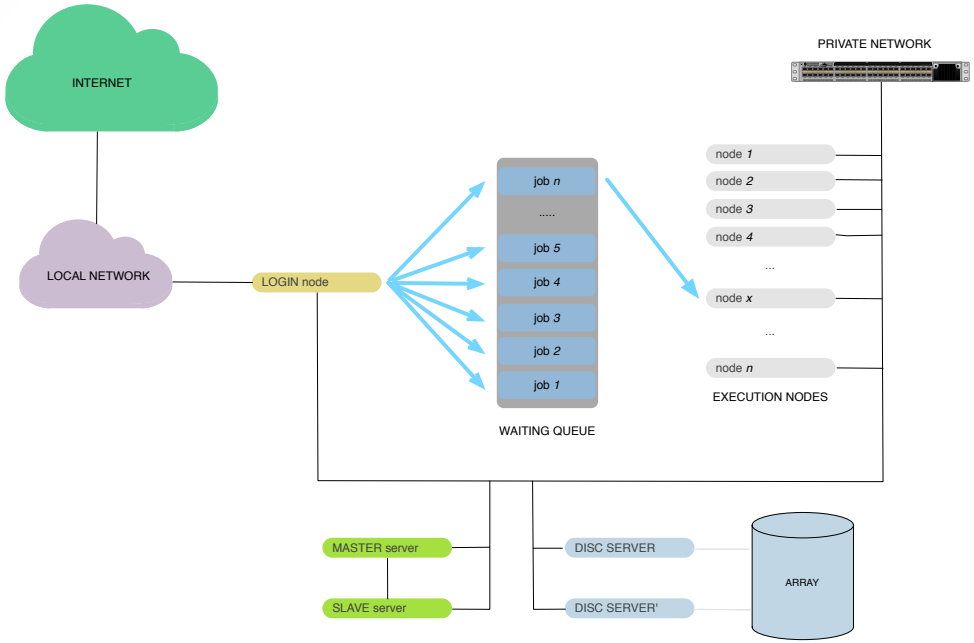
  - "*Know how*" y soporte existente (colaboradores, documentación "on-line"…)

  - Proyecto maduro (otros centros que lo tengan en producción)

  - Integración con sistemas gestores de colas (*GridEngine / Slurm*)

Nuestra experiencia III

- ***Arquitectura HPC del RDlab***



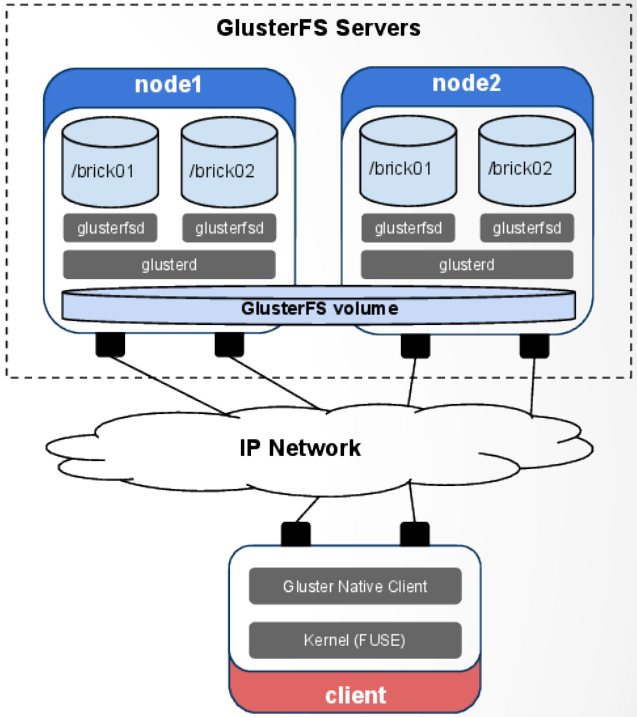
- ***Características del sistema de ficheros HPC***

  - Más de 150 millones de ficheros (directorios con +1millón de ficheros)
  - Más de 100 nodos conectados a 1 Gbit y servidores de datos a 10Gbit Ethernet
  - Entorno de pruebas con un subconjunto de unos 25Tbytes y red a 10Gbit

Nuestra experiencia IV

- ***GlusterFS (pros)***

  - Fundado en 2005 y comprado por RedHat en 2012
  - Actualmente software libre
  - Funciona sobre sistemas de ficheros **autónomos**
  - Separación de datos y metadatos (hash distribuido)
  - Tolerancia a fallos y replicación de datos
  - Fácil de instalar y requiere poca infraestructura
  - Preparado para re-exportación (CIFS/NFS)



- ***GlusterFS (contras)***

  - Recuperación del espacio libre
  - "Auto-healing"
  - Rendimiento (elección sistema de ficheros local, ficheros pequeños, interactividad)
  - Bloqueos POSIX

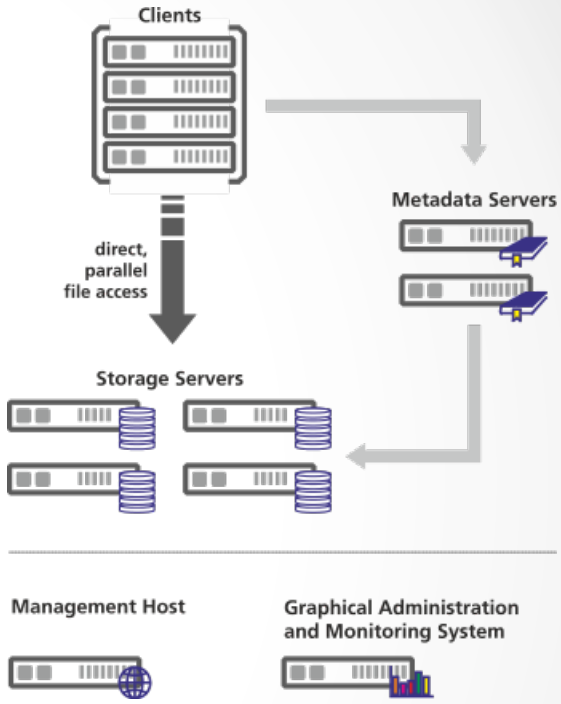
Nuestra experiencia V

- **BeeGFS/FhGFS (pros)**

  - Nacido en la universidad de Fraunhofer (2005)
  - Actualmente software libre
  - Separación **total** entre datos y metadatos
  - Fácil de instalar y requiere poca infraestructura
  - Muy buen rendimiento y escalabilidad

- **BeeGFS/FhGFS (contras)**

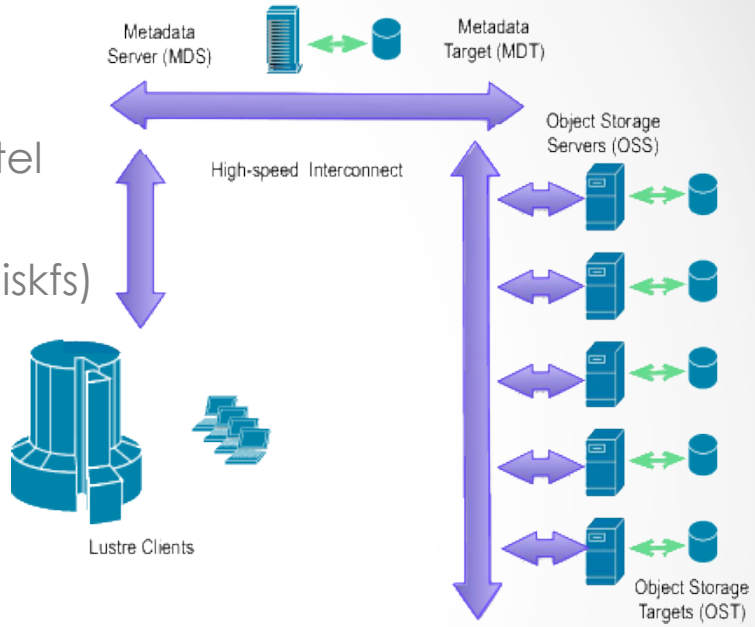
  - Proyecto poco maduro para producción
  - Poca documentación (comunidad)
  - Alta disponibilidad no nativa



Nuestra experiencia VI

- ***Lustre (pros)***

  - Entorno universitario, pasó por SUN, Oracle e Intel
  - Actualmente software libre (forks)
  - Disponibilidad de sistema de ficheros propio (ldiskfs)
  - Separación **total** entre datos y metadatos
  - Buen rendimiento y escalabilidad
  - Usado por +60% del top100 HPC
  - Bloqueos del sistema de ficheros



- ***Lustre (contras)***

  - "Difícil" de instalar (aunque existe "*patchless client*") y entender
  - Requiere de una cierta infraestructura mínima
  - Arquitectura del año 2000 (MDT no distribuido)
  - Los datos no se encuentran disponibles parcialmente en caso de fallo

Ficheros de alto rendimiento fuera del HPC

- ***Nuevos motivos***

  - Aumento exponencial del volumen de datos
  - Sistemas multicore y discos "infinitos"
  - Nuevas aplicaciones y servicios (ej. OpenData)
  - Virtualización

- ***Viejos motivos (MÁS x menos)***

  - Sistema robusto y fiable
  - Copias de seguridad y movimiento de datos (migraciones)
  - Optimización/reaprovechamiento del hardware
  - La "magia" de las TIC

Rendimiento del Lustre RDlab

- ***Benchmark I (interactividad* desde 1 y 5 clientes)***

sync && echo 1 > /proc/sys/vm/drop_caches; time gunzip proteines.tgz

Real **0m1.025s** (user 0m0.644s + sys 0m0.190s) **0m0.965s** - **0m1.017s** - **0m1.011s** - **0m1.006s** - **0m0.954s**

sync && echo 1 > /proc/sys/vm/drop_caches; time tar xf proteines.tar

Real **1m45.913s** (user 0m0.496s + sys 0m21.219s) **1m43.491s** - **1m43.231s** - **1m41.035s** - **1m43.212s** - **1m41.947s**

sync && echo 1 > /proc/sys/vm/drop_caches; time ls -la proteines-41199/ > /dev/null

Real **0m2.834s** (user 0m0.225s + sys 0m2.385s) **0m2.754s** - **0m2.700s** - **0m2.872s** - **0m3.364s** - **0m3.367s**

sync && echo 1 > /proc/sys/vm/drop_caches; time find proteines-41199/* -exec cat {} > /dev/null \;

Real **2m18.165s** (user 0m15.808s + sys 0m31.451s) **1m11.217s** - **2m16.468s** - **2m16.744s** - **2m25.126s** - **1m7.215s**

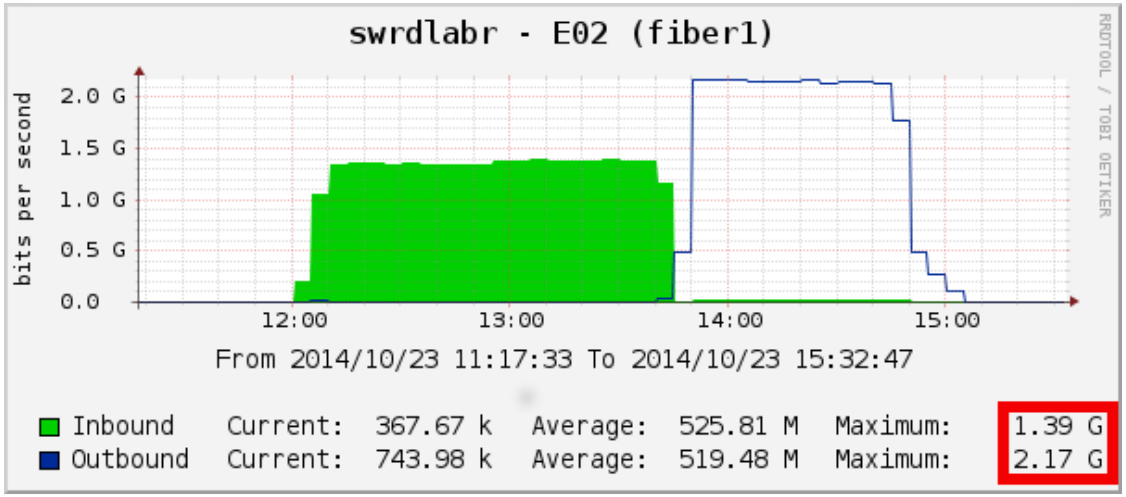
sync && echo 1 > /proc/sys/vm/drop_caches; time du -s proteines-41199/ > /dev/null

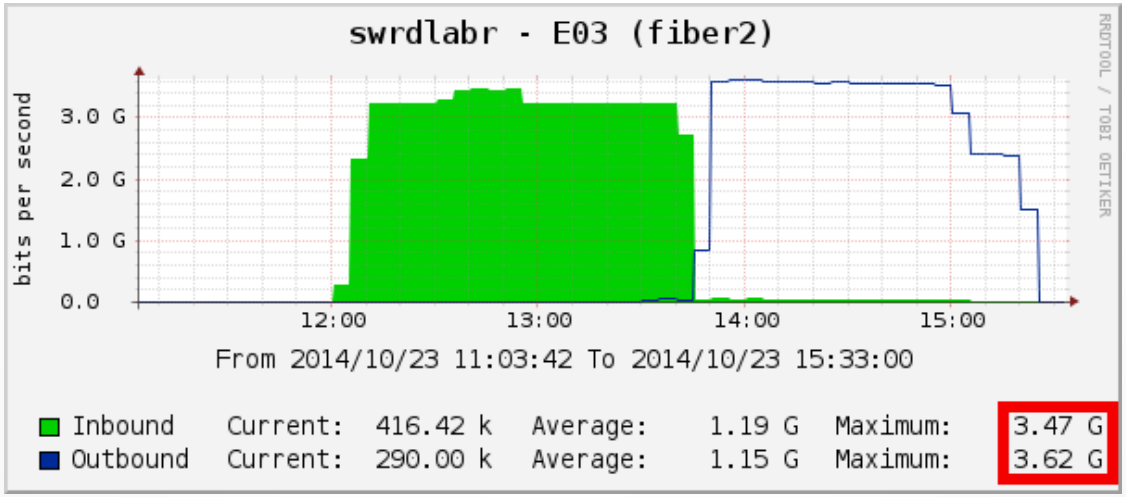
Real **0m1.683s** (user 0m0.026s + sys 0m1.597s) **0m1.726s** - **0m1.689s** - **0m1.756s** - **0m1.685s** - **0m1.760s**

(*) *proteines.tgz* es un directorio con 42k ficheros. El sistema está en producción corriendo otras aplicaciones de usuario.

Rendimiento del Lustre RDlab II

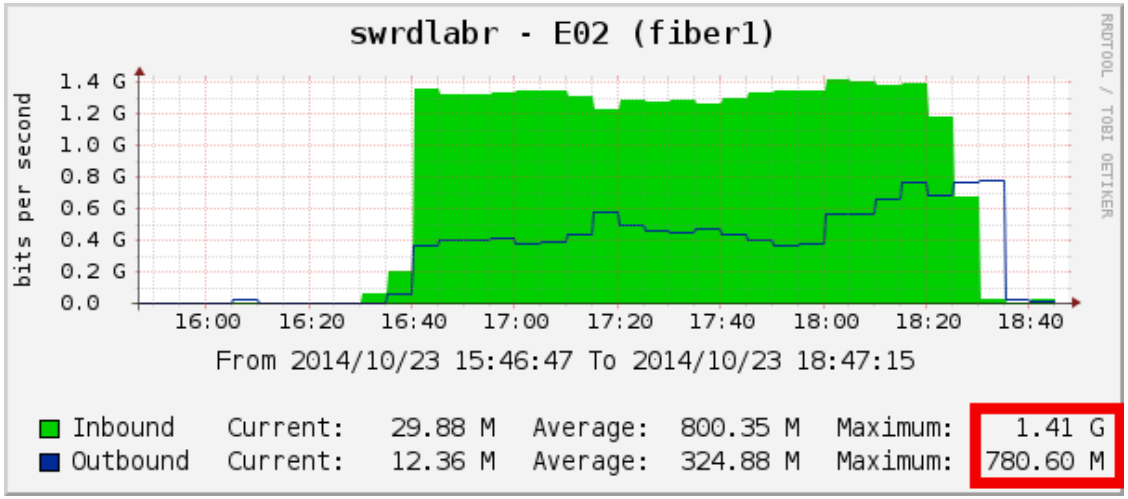
- ***Benchmark II (rendimiento sostenido con 12 clientes*)***

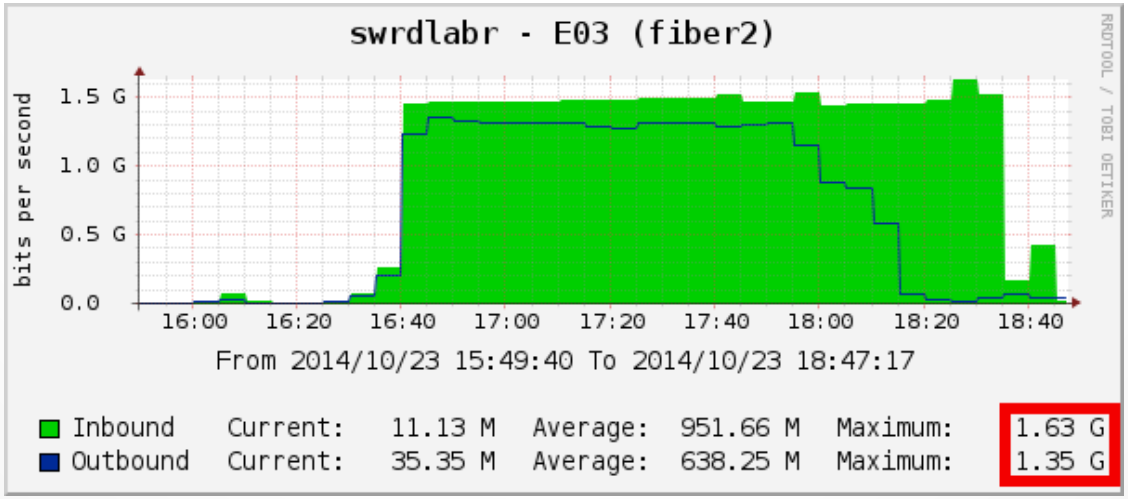




Rendimiento del Lustre RDlab III

- ***Benchmark III (rendimiento sostenido bruto con 6+6 clientes)***





¡Gracias!

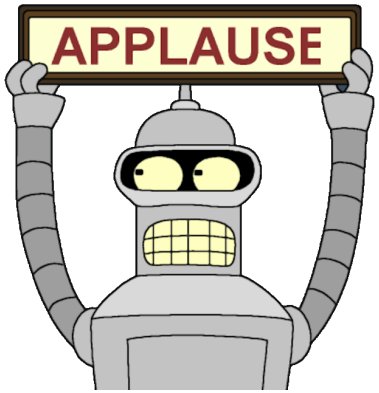
- ***Aspectos para reflexionar / preguntar…***

  - Sistemas de archivos locales vs Sistemas de archivos de alto rendimiento

  - ¿Cuáles son realmente mis necesidades?

  - No hay soluciones únicas ni universales

  - ¡Colaboremos! El RDlab quiere saber de ti.



REFERENCIAS

[www1] http://tinyurl.com/knlj2fm
[www2] http://en.wikipedia.org/wiki/Unix_File_System
[www3] http://sysadminday.com/wp-content/uploads/wppa/2.jpg?ver=1
[www4] http://www.gluster.org/
[www5] http://en.wikipedia.org/wiki/Distributed_hash_table
[www6] http://delivery.acm.org/10.1145/2560000/2555790/11549.html
[www7] http://moo.nac.uci.edu/~hjm/fhgfs_vs_gluster.html
[www8] http://en.wikipedia.org/wiki/BeeGFS
[www9] http://wiki.lustre.org/lid/ulfi/complete/ulfi_complete.html
[www10] http://en.wikipedia.org/wiki/Lustre_%28file_system%29
[www11] http://gluster.org/community/documentation/index.php/Gluster_3.2:_Accessing_Data_-_Setting_Up_GlusterFS_Client
[www12] https://wiki.hpdd.intel.com/display/PUB/HPDD+Wiki+Front+Page
[www13] http://www.fhgfs.com/

[SRS02] Linux System Administration,Vicki Stanfield, Roderick W. Smith, Ed. SYBEX, 2002